
Zusammenfassung

Die Grundkonzepte der Datenanalyse werden anhand des bekannten Iris-Datensatzes eingeführt. Datenskalen (nominal, ordinal, Intervall, proportional) müssen berücksichtigt werden, weil bestimmte mathematische Operationen nur für gewisse Skalen geeignet sind. Numerische Daten können als Mengen, Vektoren oder Matrizen repräsentiert werden. Viele Datenanalyseverfahren basieren auf Unähnlichkeitsmaßen (z. B. Matrixnormen, Lebesgue/Minkowski-Normen) oder Ähnlichkeitsmaßen (z. B. Cosinus, Überlapp, Dice, Jaccard, Tanimoto). Sequenzen können mit Sequenzrelationen analysiert werden (z. B. Hamming, Levenshtein/Edit-Abstand). Aus kontinuierlichen analogen Signalen können Daten durch Abtastung und Quantisierung extrahiert werden. Die Nyquist-Bedingung ermöglicht eine Abtastung ohne Informationsverlust.

2.1 Der Iris-Datensatz

Zur Einführung der Grundkonzepte der Datenanalyse betrachten wir einen der bekanntesten historischen Referenz-Datensätze: den *Iris-Datensatz* [1]. Der Iris-Datensatz wurde 1935 von dem amerikanischen Botaniker Edgar Anderson erstellt, der die geographischen Unterschiede der Schwertlilien (Iris) untersuchte. Die von Anderson erhobenen Daten über Schwertlilien auf der Gaspé-Halbinsel in Quebec (Kanada) wurden erstmals von Sir Ronald Aylmer Fisher 1936 als Beispiel für die Diskriminanzanalyse multivariater Daten (siehe Kap. 8) verwendet [4] und entwickelten sich später zu einem der meistverwendeten Referenzdatensätze in Statistik und Datenanalyse. Der Iris-Datensatz umfasst Messdaten von 150 Irispflanzen, davon jeweils 50 Borsten-Schwertlilien (Iris Setosa), Virginia-Schwertlilien (Iris Virginica) und verschiedenfarbige Schwertlilien (Iris Versicolor). Der Datensatz enthält von jeder dieser 150 Pflanzen die Länge und Breite eines

Kelchblatts sowie die Länge und Breite eines Blütenblatts (in Zentimetern). Den kompletten Datensatz zeigt Tab. 2.1. Inzwischen sind einige unterschiedliche Variationen dieses Datensatzes verwendet und veröffentlicht worden, so dass bei Vergleichen überprüft werden sollte, welche Variante des Iris-Datensatzes verwendet wurde [2]. Der Iris-Datensatz und viele andere bekannte Referenzdatensätze sind z. B. über die *Machine Learning Data Base* der Universität von Kalifornien (Irvine, UCI) verfügbar.

In der Datenanalyse bezeichnen wir jede der 150 Irispflanzen als ein *Objekt*, jede der 3 Pflanzenarten als eine *Klasse* und jede der 4 Messgrößen als ein *Merkmal*. Typische Fragen, die uns bei der Analyse dieser Daten interessieren, sind:

- Welche dieser Daten enthalten möglicherweise Messfehler oder falsche Klassenzuordnungen?
- Welcher Fehler wird durch die Rundung der Daten auf 0,1 Zentimeter verursacht?
- Wie stark ist der Zusammenhang zwischen Länge und Breite einer Blüte?
- Zwischen welchen Merkmalen besteht der am stärksten ausgeprägte Zusammenhang?
- Keine der Pflanzen in diesem Datensatz hat eine Kelchbreite von 1,8 Zentimetern. Welche Kelchlänge würden wir bei einer solchen Pflanze erwarten?
- Zu welcher Pflanzenart gehört vermutlich eine Iris mit einer Kelchbreite von 1,8 Zentimetern?
- Sind in den drei betrachteten Arten auch Unterarten enthalten, die aus den Daten ersichtlich werden?

In diesem Buch werden wir zahlreiche Methoden kennenlernen, solche und ähnliche Fragestellungen zu beantworten. Zum Verständnis dieser Methoden ist es notwendig, zunächst einige grundlegende Eigenschaften von Daten und deren Relationen zueinander formal zu definieren.

2.2 Maßskalen

Numerische Informationen können unterschiedliche Bedeutung haben, auch wenn sie durch die gleichen numerischen Daten repräsentiert werden. Je nach semantischer Bedeutung können bestimmte mathematische Operationen zulässig oder unzulässig sein. Für die semantische Bedeutung numerischer Daten wurden von Stevens [7] vier unterschiedliche Skalen vorgeschlagen (Tab. 2.2). Für nominalskalierte Daten (unterste Zeile) sind nur Test auf Gleichheit und Ungleichheit zulässig. Beispiele für nominalskalierte Daten sind Namen von Personen oder Indizes von Objekten. Daten eines nominalskalierten Merkmals können durch den *Modus* (oder *Modalwert*), also dem am häufigsten vorkommenden Wert, repräsentiert werden. Für ordinalskalierte Daten (zweite Zeile von unten) sind die Ordnungsrelationen $>$ bzw. $<$ gültig. Auf jedem Skalen-Niveau sind auch die Operationen und statistischen Maße aller niedrigen Skalen-Niveaus gültig. Für ordinalskalierte Daten

Tab. 2.1 Iris-Datensatz [1]

Setosa				Versicolor				Virginica			
Kelchblatt		Blütenblatt		Kelchblatt		Blütenblatt		Kelchblatt		Blütenblatt	
Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.4	3.3	1	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3	1.6	0.2	6.6	3	4.4	1.4	7.2	3.2	6	1.8
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3	5	1.7	6.1	3	4.9	1.8
5.2	3.4	1.4	0.2	6	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1	7.2	3	5.8	1.6

Tab. 2.1 (Fortsetzung)

Setosa				Versicolor				Virginica			
Kelchblatt		Blütenblatt		Kelchblatt		Blütenblatt		Kelchblatt		Blütenblatt	
Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1	7.9	3.8	6.4	2
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3	4.5	1.5	6.1	2.6	5.6	1.4
5	3.2	1.2	0.2	6	3.4	4.5	1.6	7.7	3	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Tab. 2.2 Maßskalen

Skala	Operation		Beispiel	Statistisches Maß
Proportional	·	/	273° K, 21 Jahre	Verallgemeinerter Mittelwert
Intervall	+	−	20° C, 2020 n. Chr.	Mittelwert
Ordinal	>	<	Sehr gut, gut, befriedigend	Median
Nominal	=	≠	Müller, Meier, Schulz	Modus

sind also die Operationen $=$, \neq , $>$ und $<$ gültig, und somit auch die Kombinationen „größer oder gleich“ (\geq) und „kleiner oder gleich“ (\leq). Die Relation „kleiner oder gleich“ (\leq) definiert eine *totale Ordnung*, so dass für alle x, y, z gilt $(x \leq y) \wedge (y \leq x) \Rightarrow (x = y)$ (Antisymmetrie), $(x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z)$ (Transitivität) und $(x \leq y) \vee (y \leq x)$

(Totalität). Beispiele für ordinalskalierte Daten sind Schulnoten. Daten eines ordinalskalierten Merkmals können durch den *Median* repräsentiert werden, also den Wert für den (ungefähr) so viele größere wie kleinere Werte vorliegen. Der *Mittelwert* ist für ordinalskalierte Daten nicht zulässig. Es ist also nicht sinnvoll, von einem Noten-Mittelwert von 3,0 zu sprechen. Für intervallskalierte Daten (dritte Zeile von unten) sind Addition und Subtraktion gültig. Intervallskalierte Merkmale haben kontextspezifisch definierte Nullpunkte. Beispiele sind Jahreszahlen nach Christus oder Temperaturen in Grad Celsius oder Grad Fahrenheit. Es ist also nicht sinnvoll zu sagen, dass 40°C doppelt so warm ist wie 20°C . Daten eines intervallskalierten Merkmals, z. B. die Daten $X = \{x_1, \dots, x_n\}$, können durch den (arithmetischen) *Mittelwert*

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \quad (2.1)$$

repräsentiert werden. Für proportionalskalierte Daten (oberste Zeile) sind Multiplikation und Division gültig. Beispiele für proportionalskalierte Merkmale sind Zeitdifferenzen (z. B. Alter) oder Temperaturen auf der Kelvin-Skala. Daten eines proportionalskalierten Merkmals können durch den *verallgemeinerten Mittelwert*

$$m_\alpha(X) = \sqrt[\alpha]{\frac{1}{n} \sum_{k=1}^n x_k^\alpha} \quad (2.2)$$

$\alpha \in \mathbb{R}$, repräsentiert werden. Der verallgemeinerte Mittelwert enthält die Sonderfälle Minimum ($\alpha \rightarrow -\infty$), harmonischer Mittelwert ($\alpha = -1$), geometrischer Mittelwert ($\alpha \rightarrow 0$), arithmetischer Mittelwert ($\alpha = 1$), quadratischer Mittelwert ($\alpha = 2$) und Maximum ($\alpha \rightarrow \infty$).

Die Messdaten aus dem Iris-Datensatz sind proportionalskalierte Daten. So kann zum Beispiel aus der Länge und Breite von Kelch- und Blütenblättern durch Multiplikation näherungsweise die Fläche der Kelch- und Blütenblätter bestimmt werden. Es sind daher alle statistischen Maße Modus, Median, Mittelwert und verallgemeinerter Mittelwert zulässig. Wir berechnen diese für die Blütenblattbreite (viertes Merkmal). Der Datensatz enthält Blütenblattbreiten zwischen 0,1 und 2,5 Zentimetern. Die am häufigsten vorkommende Blütenblattbreite ist 0,2 Zentimeter, und zwar bei 29 Pflanzen. Der Modalwert der Blütenblattbreite ist also 0,2 Zentimeter. Zur Berechnung des Median wird die Anzahl der Pflanzen mit den Blütenblattbreiten 0,1 Zentimeter, 0,2 Zentimeter, 0,3 Zentimeter usw. aufsummiert, bis die Hälfte der Objekte (75) erfasst ist. Dies ist im linken Teil von Tab. 2.3 dargestellt. Der Median der Blütenblattbreite ist also 1,3 Zentimeter. Alternativ kann die Summierung auch in absteigender Reihenfolge (rechter Teil von Tab. 2.3) erfolgen. Dieser Algorithmus hat die zeitliche Komplexität $O(n \log n)$. Allerdings (und das ist selbst in der Wissenschaft wenig bekannt) kann der Median mit Hilfe sogenannter Selektionsalgorithmen effizient in linearer Zeit berechnet werden [3]. Der Mittelwert der Blütenblattbreite ist schließlich ca. 1,19933.

Viele Methoden, die in diesem Buch beschrieben sind, gehen von intervall- oder proportionalskalierten Daten aus und verwenden das entsprechende maßtheoretisch zulässige

Tab. 2.3 Berechnung des Medians der Blütenblattbreite (Iris)

Wert	Anzahl	kumulierte Summe	Wert	Anzahl	kumulierte Summe
0,1	5	5	2,5	3	3
0,2	29	34	2,4	3	6
0,3	7	41	2,3	8	14
0,4	7	48	2,2	3	17
0,5	1	49	2,1	6	23
0,6	1	50	2	6	29
0,7	0	50	1,9	5	34
0,8	0	50	1,8	12	46
0,9	0	50	1,7	2	48
1	7	57	1,6	4	52
1,1	3	60	1,5	12	64
1,2	5	65	1,4	8	72
1,3	10(13)	75	1,3	3(13)	75

mathematische Instrumentarium. Es werden jedoch auch zahlreiche Methoden zur Analyse von nominal- und ordinalskalierten Daten beschrieben. Diese erfolgt meist mit Hilfe von Relationen bzw. relationalen Analysemethoden.

2.3 Mengen- und Matrixdarstellung

Jeder numerischer Merkmalsdatensatz lässt sich als Menge

$$X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p \tag{2.3}$$

schreiben. Ein solcher Datensatz enthält $n \in \{1, 2, \dots\}$ Elemente. Jedes Element ist ein p -dimensionaler reellwertiger Merkmalsvektor, $p \in \{1, 2, \dots\}$. Für $p = 1$ heißt X *skalarer Datensatz*. Außer der Mengenschreibweise wird häufig auch die Matrixschreibweise

$$X = \begin{pmatrix} x_1^{(1)} & \cdots & x_1^{(p)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(p)} \end{pmatrix} \tag{2.4}$$

verwendet. Die Vektoren x_1, \dots, x_n sind also *Zeilenvektoren*. Mathematisch nicht ganz sauber werden die Menge X und die Matrix X häufig als äquivalente Repräsentationen eines Datensatzes benutzt. Abbildung 2.1 zeigt die üblichen Bezeichnungen der Elemente von Datenmatrizen. Jede *Zeile* in der Datenmatrix entspricht einem *Element* der Datenmenge und wird als *Merkmalsvektor* oder *Datenpunkt* x_k bezeichnet, $k = 1, \dots, n$. Jede

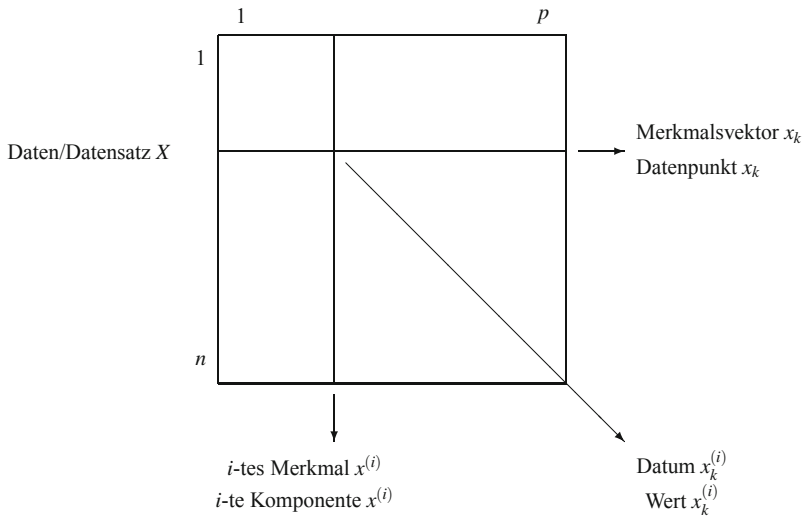


Abb. 2.1 Matrixschreibweise eines Datensatzes

Spalte in der Datenmatrix entspricht einer *Komponente* aller Elemente der Datenmenge und wird als *i*-tes *Merkmal* oder *i*-te *Komponente* $x^{(i)}$ bezeichnet, $i = 1, \dots, p$. Zeilen und Spalten werden in diesem Buch durch tiefgestellte Indizes für Zeilen und eingeklammerte hochgestellte Indizes für Spalten unterschieden. In der Literatur finden sich auch alternative Schreibweisen, zum Beispiel $x(k, \cdot)$ und $x(\cdot, i)$. Ein einzelnes *Matrixelement* entspricht *einer* Komponente *eines* Elements der Datenmenge und wird *Datum* oder *Wert* $x_k^{(i)}$ genannt, $k = 1, \dots, n, i = 1, \dots, p$.

Der Iris-Datensatz lässt sich als eine solche Datenmatrix mit 150 Zeilen und 4 Spalten darstellen. Jede Zeile dieser Matrix entspricht einer der 150 Pflanzen, und jede Spalte der Matrix entspricht einer der 4 Messgrößen. Die Iris-Datenmatrix kann also durch Untereinanderhängen der drei Teile in Tab. 2.1 erzeugt werden. Die Klasseninformation (Setosa, Versicolor, Virginica) kann als zusätzliches fünftes (ordinalskaliertes) Merkmal in einer solchen Datenmatrix interpretiert werden.

2.4 Relationen

Ohne Bezug auf numerische Merkmale schreiben wir eine Menge von (abstrakten) Elementen als

$$O = \{o_1, \dots, o_n\} \quad (2.5)$$

Häufig lassen sich solche Objekte $o_k, k = 1, \dots, n$, nicht sinnvoll mit Merkmalsvektoren repräsentieren, so dass konventionelle merkmalsbasierte Datenanalysemethoden nicht verwendet werden können. Stattdessen lässt sich oft eine *Relation* aller Paare von Objekten

quantifizieren und als quadratische Matrix

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n} \quad (2.6)$$

schreiben. Durch die Relationswerte r_{ij} , $i, j = 1, \dots, n$, lassen sich Grade der Ähnlichkeit, Unähnlichkeit, Vereinbarkeit, Unvereinbarkeit, Nähe oder Abstand zwischen den Objekten o_i und o_j quantifizieren. R heißt *symmetrisch*, wenn $r_{ij} = r_{ji}$ für alle $i, j = 1, \dots, n$. Die Werte in R können manuell definiert oder aus Merkmalen berechnet werden. Wenn sinnvolle numerische Merkmale X verfügbar sind, dann kann R durch eine geeignete Funktion $f: \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ berechnet werden. Zum Beispiel könnte eine Relationsmatrix für die Iris-Daten manuell durch einen Botaniker spezifiziert werden, der die Ähnlichkeit jedes Paares von Pflanzen begutachtet und dann numerisch quantifiziert, oder sie könnte mit einer geeigneten Funktion f aus den Längen und Breiten der Kelch- und Blütenblätter berechnet werden. Die folgenden beiden Abschnitte behandeln die in der Datenanalyse wichtigsten Klassen von Relationen: Unähnlichkeiten und Ähnlichkeiten.

2.5 Unähnlichkeitsmaße

Eine Funktion d heißt *Unähnlichkeitsmaß* oder *Distanzmaß* wenn für alle $x, y \in \mathbb{R}^p$ gilt

$$d(x, y) = d(y, x) \quad (2.7)$$

$$d(x, y) = 0 \Leftrightarrow x = y \quad (2.8)$$

$$d(x, z) \leq d(x, y) + d(y, z) \quad (2.9)$$

Aus diesen Axiomen folgt

$$d(x, y) \geq 0 \quad (2.10)$$

Eine Klasse von Unähnlichkeitsmaßen ist definiert durch eine *Norm* $\|\cdot\|$ von $x - y$, also

$$d(x, y) = \|x - y\| \quad (2.11)$$

Eine Funktion $\|\cdot\|: \mathbb{R}^p \rightarrow \mathbb{R}^+$ heißt *Norm* genau dann, wenn

$$\|x\| = 0 \Leftrightarrow x = (0, \dots, 0) \quad (2.12)$$

$$\|a \cdot x\| = |a| \cdot \|x\| \quad \forall a \in \mathbb{R}, x \in \mathbb{R}^p \quad (2.13)$$

$$\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^p \quad (2.14)$$

Zum Beispiel ist die häufig benutzte sogenannte *hyperbolische Norm*

$$\|x\|_h = \prod_{i=1}^p x^{(i)} \quad (2.15)$$

keine Norm gemäß obiger Definition, denn die Bedingung (2.12) wird verletzt durch $x = (0, 1) \neq (0, 0)$ mit $\|x\|_h = \|(0, 1)\|_h = 0$, bzw. Bedingung (2.13) wird verletzt durch $x = (1, 1)$ und $a = 2$ mit $\|a \cdot x\|_h = \|2 \cdot (1, 1)\|_h = \|(2, 2)\|_h = 4 \neq |a| \cdot \|x\|_h = |2| \cdot \|(1, 1)\|_h = 2$.

Zu den häufig verwendeten Normen gehören die *Matrixnormen* und die Lebesgue- oder *Minkowski-Normen*. Die Matrixnorm ist definiert als

$$\|x\|_A = \sqrt{x A x^T} \quad (2.16)$$

mit einer Matrix $A \in \mathbb{R}^{n \times n}$. Wichtige Sonderfälle von Matrixnormen sind die *Euklidische Norm*

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \quad (2.17)$$

die *Frobenius- oder Hilbert-Schmidt-Norm*

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \quad (2.18)$$

die *Diagonálnorm* mit merkmalspezifischen Gewichten

$$A = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix} \quad (2.19)$$

und die *Mahalanobis-Norm*

$$A = \text{cov}^{-1} X = \left(\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^T (x_k - \bar{x}) \right)^{-1} \quad (2.20)$$

Die Mahalanobis-Norm verwendet die Inverse der Kovarianzmatrix des Datensatzes X (siehe Kap. 5). Sie passt die Gewichtung der einzelnen Komponenten an die beobachtete Statistik an und berücksichtigt auch Korrelationen zwischen Merkmalen.

Die Lebesgue- oder Minkowski-Norm ist definiert als

$$\|x\|_\alpha = \sqrt[\alpha]{\sum_{j=1}^p |x^{(j)}|^\alpha} \quad (2.21)$$

und entspricht dem verallgemeinerten Mittelwert (2.2) bis auf einen konstanten Faktor $\sqrt[p]{n}$. Wichtige Sonderfälle der Lebesgue- oder Minkowski-Norm sind die Infimum-Norm ($\alpha \rightarrow -\infty$)

$$\|x\|_{-\infty} = \min_{j=1,\dots,p} x^{(j)} \quad (2.22)$$

die *Manhattan-* oder *City-Block-Norm* ($\alpha = 1$)

$$\|x\|_1 = \sum_{j=1}^p |x^{(j)}| \quad (2.23)$$

die *Euklidische Norm* ($\alpha = 2$) als einziger Schnittpunkt zwischen Matrixnormen und Lebesgue-/Minkowski-Normen

$$\|x\|_2 = \sqrt{\sum_{j=1}^p (x^{(j)})^2} \quad (2.24)$$

und die Supremum-Norm ($\alpha \rightarrow \infty$)

$$\|x\|_{\infty} = \max_{j=1,\dots,p} |x^{(j)}| \quad (2.25)$$

Ein weiteres häufig verwendetes Unähnlichkeitsmaß ist der *Hamming-Abstand* [5]

$$d_H(x, y) = \sum_{i=1}^p \rho(x^{(i)}, y^{(i)}) \quad (2.26)$$

mit der diskreten Metrik

$$\rho(x, y) = \begin{cases} 0 & \text{falls } x = y \\ 1 & \text{sonst} \end{cases} \quad (2.27)$$

Der Hamming-Abstand ist also die Anzahl der unterschiedlichen Merkmalswerte. Für binärwertige Merkmale ist der Hamming-Abstand gleich dem Manhattan- oder City-Block-Abstand (2.23), $d_H(x, y) = \|x - y\|_1$. Der Hamming-Abstand ist jedoch keine Norm, denn Bedingung (2.13) ist verletzt. Eine Variante des Hamming-Abstands ersetzt (2.27) durch andere Funktionen ρ , um die Ähnlichkeit zwischen Merkmalen zu quantifizieren. Sind die Merkmale beispielsweise (nominalskalierte) Teile einer Maschine, dann könnten ähnliche Teile zu geringeren Werten von ρ führen als weniger ähnliche Teile.

2.6 Ähnlichkeitsmaße

Eine Funktion s heißt *Ähnlichkeitsmaß* wenn für alle $x, y \in \mathbb{R}^p$ gilt

$$s(x, y) = s(y, x) \quad (2.28)$$

$$s(x, y) \leq s(x, x) \quad (2.29)$$

$$s(x, y) \geq 0 \quad (2.30)$$

Die Funktion s heißt *normalisiertes Ähnlichkeitsmaß*, wenn zusätzlich gilt

$$s(x, x) = 1 \quad (2.31)$$

Jedes Unähnlichkeitsmaß d kann genutzt werden, um ein Ähnlichkeitsmaß s zu definieren (und umgekehrt), etwa mit einer monoton fallenden positiven Funktion f mit $f(0) = 1$ wie beispielsweise

$$s(x, y) = \frac{1}{1 + d(x, y)} \quad (2.32)$$

Die Beispiele aus dem vorherigen Abschnitt werden jedoch meist als Unähnlichkeit d verwendet und die Beispiele in diesem Abschnitt meist als Ähnlichkeit s .

Betrachten wir zunächst Ähnlichkeiten zwischen *binären* Merkmalsvektoren. Zwei binäre Merkmalsvektoren können als ähnlich betrachtet werden, wenn viele ihrer Einsen übereinstimmen. Die Anzahl der übereinstimmenden Einsen kann mit dem Skalarprodukt berechnet werden. Bei der Verallgemeinerung für beliebige positive reellwertige Merkmale werden Ähnlichkeitsmaße als unterschiedlich normalisierte Skalarprodukte definiert:

- Kosinus

$$s(x, y) = \frac{\sum_{i=1}^p x^{(i)} y^{(i)}}{\sqrt{\sum_{i=1}^p (x^{(i)})^2 \sum_{i=1}^p (y^{(i)})^2}} \quad (2.33)$$

- Überlapp

$$s(x, y) = \frac{\sum_{i=1}^p x^{(i)} y^{(i)}}{\min\left(\sum_{i=1}^p (x^{(i)})^2, \sum_{i=1}^p (y^{(i)})^2\right)} \quad (2.34)$$

- Dice

$$s(x, y) = \frac{2 \sum_{i=1}^p x^{(i)} y^{(i)}}{\sum_{i=1}^p (x^{(i)})^2 + \sum_{i=1}^p (y^{(i)})^2} \quad (2.35)$$

- Jaccard (auch Tanimoto)

$$s(x, y) = \frac{\sum_{i=1}^p x^{(i)} y^{(i)}}{\sum_{i=1}^p (x^{(i)})^2 + \sum_{i=1}^p (y^{(i)})^2 - \sum_{i=1}^p x^{(i)} y^{(i)}} \quad (2.36)$$

Data Mining

Modelle und Algorithmen intelligenter Datenanalyse

Runkler, Th.A.

2015, XII, 145 S. 72 Abb., Softcover

ISBN: 978-3-8348-1694-8